基于 YOLO-IDOD 的红外动态目标实时检测算法

赵　鑫[1,2], 费晓虎[1], 王东宇[1], 韩守飞[1]

(1.安徽理工大学 人工智能学院,安徽 淮南 232001;2.安徽理工大学 煤炭无人化开采数智技术全国重点实验室,安徽 淮南 232001)

摘　要:针对现有红外目标检测算法对动态目标检测时存在未充分利用时序信息挖掘连续帧之间的关联性导致检测精度不高的问题,提出一种以 DAM 与 CACONV 为核心的基于 YOLO-IDOD 的红外动态目标实时检测算法。以 YOLOv12s 作为基础网络架构,首先,在输入端引入动态关注模块,使用光流网络计算短时光流特征,抑制背景运动噪声,使网络关注实际目标的运动特征,提升检测精度;其次,在网络架构中引入通道注意力卷积模块,该模块在输入通道与输出通道均增加通道注意力机制,使网络能够更好地理解与关注 DAM 模块输入的数据特征;最后,将上述模块作为优化动态目标检测模型即插即用模块,使网络具备时空聚合与特征选择能力,提升网络对于红外动态目标检测的泛化性能。实验结果表明:改进后的 YOLO-IDOD 模型在自建数据集 IRDA 和公共数据集 FLIR_ADAS_v2 的混合数据集上对红外动态目标检测取得的准确率 P、召回率 R、$mAP@50$ 和 $mAP@95$ 分别为 79.9%、62.5%、77.7%和 57.3%,相较于改进前的 YOLOv12s 基准模型在维持召回率的同时准确率提升 5.2 百分点、$mAP@50$ 提升 4.6 百分点、$mAP@95$ 提升 2.4 百分点,有效提升了对于动态目标的检测精度与泛化能力。

关键词:红外动态目标检测; YOLOv12; DAM; CACONV; 多维通道注意力机制

中图分类号:TP391.41;TN219　　**文献标志码**:A　　**doi**:10.13705/j.issn.1671-6833.2026.05.001

红外成像相较于可见光成像可以保存更多的信息,有较强的穿透力,具有不受恶劣环境和光照的影响等特点[1]。随着产业的升级发展,动态目标成为红外图像中的主要检测目标类别,该类目标在时间轴上具有重要特征信息。然而红外图像中的运动目标由于缺乏颜色信息、背景噪声复杂、具有较低的温度对比度,导致成像模糊[2],检测精度低于静态目标。如何在保证实时性的基础上,提高红外动态目标检测精度是当前红外目标检测领域中的关键[3]。

近年来,将注意力机制[4]与各轻量级目标检测网络[5]进行融合从而实现更为优异的红外图像下特定对象检测性能是红外目标检测中的热门研究领域。郭浩帆等[6]通过将 YOLO(you only look once)与通道和空间注意力机制进行融合从而提升对红外图像下气体的检测精度。Dai 等[7]提出了一种模型驱动的深度网络(attentional local contrast network, ALCNet),该网络将判别网络与传统的模型驱动方法相结合从而实现对于小目标的检测。Yue 等[8]提出了一种融合超分辨率技术和与多尺度观测机制结合的红外小目标检测模型(YOLO-MST),该模型通过结合图像超分辨率技术和优化骨干网络结构提升检测精度。Wang 等[9]通过在 c2f 模块中引入膨胀残差模块(dilation-wise residual, DWR),采用内容引导注意力机制进行特征融合,实现对红外图像下车辆和行人的定向检测增强。Sun 等[10]提出了一种道路红外目标检测的 YOLO 检测模型(infrared-YOLO,I-YOLO),通过增强特征提取降低红外图像噪声提高检测性能。Ling 等[11]通过对 YOLOv8[12]进行改进,利用多尺度注意特征融合的思想对算法颈部部分进行重构,优化网络结构,增强多尺度特征融合性能,实现对红外图像下道路目标的检测能力。Wang 等[13]提出一种基于 YOLOv5s[12]的改进检测模型,基于融合注意力机制的特征融合模块增强网络的特征融合,实现了对红外图像的近岸红外船舶检测精度提高。Zhao 等[14]通过优化 YOLOv8 的主干网络、替换固定卷积核等算法,在基础网络上提出

收稿日期:2025-11-12;**修订日期**:2026-01-28

基金项目:国家自然科学基金资助项目(62306279);安徽省自然科学基金资助项目(2208085ME128)

作者简介:赵鑫(1991—),男,山西运城人,安徽理工大学讲师,博士,主要从事红外图像处理与分割、语义融合研究,E-mail:zhaoxin@aust.edu.cn。

一种无人机红外目标检测模型(ITD-YOLOv8),降低了漏检和误检率。Hao 等[15]针对红外遥感对象在复杂背景和低对比度条件下缺乏对噪声和小尺寸目标的鲁棒性问题,构建深度学习网络(YOLO-Super Resolution, YOLO-SR),通过在骨干层池化模块后引入瓶颈 Transformer 模块捕捉图像中的长距离依赖关系,在颈部层设计并引入 C3-Neck 模块融合空间中时空信息,有效解决了漏检和误报问题。

上述研究使用网络主要由基础网络进行衍生,此类网络输入端普遍采用单帧图像作为输入图像,通过引入空间注意力、通道注意力或特征重构模块等模块,以提升模型对图像中低对比度、低纹理目标的辨识能力。但此类方法[6-14]通常未利用时序特征信息,忽视动态红外场景中时间信息的建模能力,难以充分挖掘连续帧之间的时序关联性,导致对动态目标检测精度低于静态目标。

YOLOv12[16]作为 YOLO 系列的最新版本,在结构设计和性能表现上均实现了显著提升,有效增强了特征提取与多尺度信息融合能力,使其在保持高速推理的同时,具备更强的目标感知与背景抑制能力。但其原有结构采用的是固定权重的卷积处理方式,对所有通道特征一视同仁,这在动态特征明显的红外图像检测任务中难以区分动态特征与背景干扰,限制了网络对动态目标的精准识别能力。

为解决现有方法中未进行时空信息融合和对泛化动态目标强化检测的缺陷,提升红外图像下动态目标的检测精度,本文提出了一种以动态关注模块(dynamic attention module, DAM)与通道注意力卷积模块(channel attention convolution, CACONV)为核心的基于 YOLO-IDOD(you only look once-infrared dynamic object detection)的红外动态目标实时检测算法,主要创新如下:

(1)在输入端接入动态关注模块 DAM,强化输入端动态目标的动态特征。

(2)在网络架构中将部分 1×1 卷积块替换为通道注意力卷积模块 CACONV,使网络能够更好地关注包含 DAM 模块输入的动态特征。

(3)对 YOLOv12[16]网络针对红外动态目标检测进行对应结构优化,有效地提升了模型对红外图像动态目标的检测精度和泛化能力,提升网络检测精度。

1 YOLOv12 模型

2015 年,Redmon 等[17]提出第一个真正意义上实现端到端单阶段高效目标检测的网络架构 YOLO,开创性地提出将目标检测转换为回归任务,大幅度提升了目标检测的实时性,当前 YOLO 系列网络仍是目标检测领域最具代表性、影响力最为广泛的 One-Stage 目标检测框架[18]与 baseline[19]。而 YOLO 系列最新的 YOLOv12[16]模型代表了实时目标检测领域的一次重大突破,标志着 YOLO 系列首次完全摆脱了传统卷积神经网络(CNN)[20]的约束,将注意力机制直接融入目标检测框架的核心设计中,其网络结构如图 1 所示。

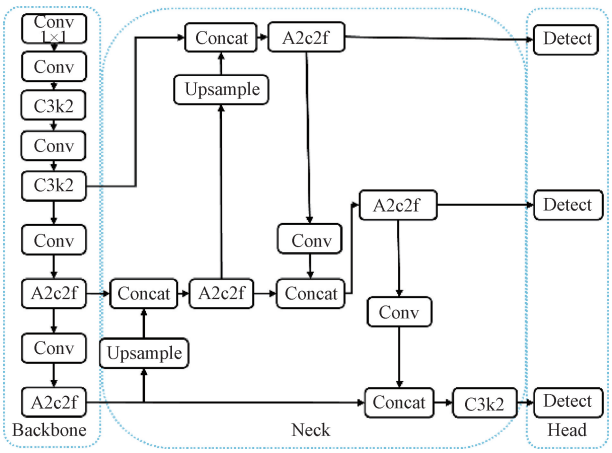


图 1 YOLOv12 网络架构

Figure 1 YOLOv12 network architecture

一方面,YOLOv12 在网络结构中引入残差高效层聚合网络[16](residual efficient layer aggregation network, R-ELAN),其结构如图 2 所示,该模块用以解决注意力机制[4]带来的优化难题,同时提出局部注意力机制用以保持高效性的前提下提升模型对关键区域的关注能力,高效层聚合网络使 YOLOv12[16]成为 YOLO 系列中首个以注意力为中心的架构,其不同于以往系列模型中使用基于 CNN[20]的传统方法的同时依然保持实时推理能力。

另一方面,YOLOv12[16]通过引入区域注意力机制(area attention, A^2)将特征图划分为相等区域来降低计算复杂度,同时保留大感受野。相比传统自注意力机制的 $O(n^2)$复杂度,区域注意力将计算成本降低 50%。区域注意力机制由此成为 YOLOv12 的核心创新,通过降低计算复杂度与提高内存访问效率解决了传统自注意力[21]的两大瓶颈。

同时 YOLOv12 包含多项精细化的架构改进,如 MLP 比例调整、位置编码替代、减少堆叠块的数量、卷积算子集成等[16],这种设计使架构更简洁,速度更快。其在与 RT-DETR[22]系列的基准比较中,YOLOv12-S 比 RT-DETRv2-R18[23]快 42%,计算量和参数分别减少 36%与 45%。

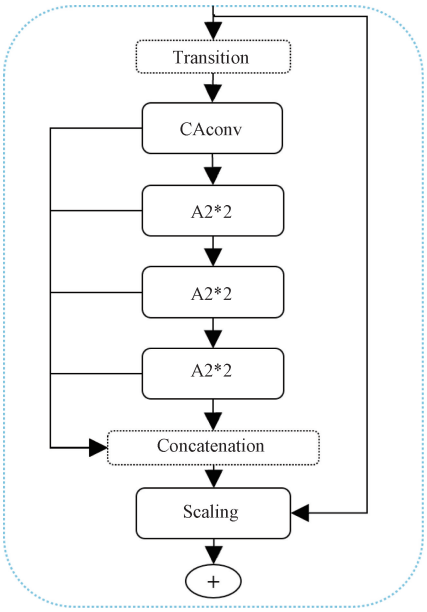


图 2　R-ELAN[16] 架构

Figure 2　R-ELAN[16] architecture

2　模型改进

2.1　动态关注模块 DAM

在红外目标检测任务中,动态目标受到运动模糊的影响从而导致目标边界与背景混淆,往往只能提供较少的特征信息。而主流的目标检测方法仍基于单帧图像进行检测与识别,最终导致基于单帧红外图像的动态目标检测算法误检率高、漏检严重。为了解决这一问题,DAM 利用光流信息提取目标的运动特征,并将目标的运动信息以光流的形式与原始图像进行通道融合,从而在输入端提高模型的时空间特征提取能力。

DAM 模块结构如图 3 所示,输入端由时间线上的连续帧组成,模块内部将连续帧分为前向运动对和后向运动对,并分别将前向运动对和后向运动对送入光流特征提取模块。该模块由 2 个相同的 LiteFlowNet v2[24] 网络组成,这 2 个光流网络分别从前向运动对和后向运动对提取当前图像对内的动态对象在时间上前向与后向光流信息,如式(1)、(2)[24] 所示。最后将蕴含运动关系的光流信息与原始图像进行通道融合作为模块输出端结果,如式(3)所示。

$$F_1 = \text{LiteFlowNetV2}(I_{t-3}, I_t); \tag{1}$$

$$F_2 = \text{LiteFlowNetV2}(I_t, I_{t+3}); \tag{2}$$

$$F = \text{Concat}(F_1, F_2, I_t)。 \tag{3}$$

式中:I_t 表示 t 时刻下的视频帧;Concat 运算表示三通道融合;F 为通道融合输出图像,包含前向运动信息与后向运动信息。

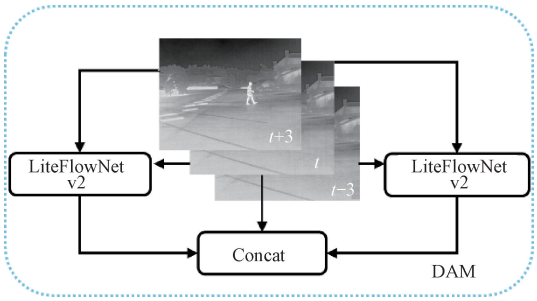


图 3　DAM 结构

Figure 3　DAM structure

2.2　通道注意力卷积模块 CACONV

CACONV 可以学习调整各输入通道的权重比例,使模型能够更加关注关键特征通道,抑制冗余通道造成的信息干扰,图 4 所示 CACONV 结构由 3 部分组成。

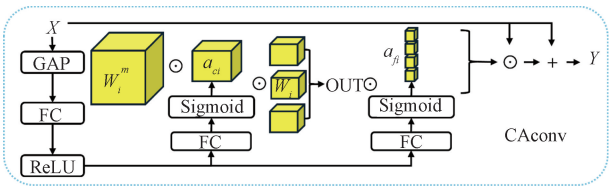


图 4　CACONV 结构

Figure 4　CACONV structure

(1)通道信息提取。通过对输入特征图 F(形状:$[B,C,H,W]$,其中 B 为批量数,C 为通道数,H 为纵向像素数,W 为横向像素数)进行全局平均池化(global average pooling, GAP)[25] 获取每个通道的全局信息,如式(4)[25] 所示:

$$S_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j)。 \tag{4}$$

式中:S_c 为 c 个通道的平均值;$F_c(i,j)$ 代表第 c 个通道在 (i,j) 的像素值;H、W 代表特征图的高与宽。这一步去除了空间信息,只保留通道信息,使得后续计算能专注于通道级别的优化。

(2)通道权重计算。通道权重计算公式如式(5)[26] 所示,采用全连接+非线性激活函数计算每个通道的重要性。

$$\boldsymbol{w} = \sigma(\boldsymbol{w}_2 \cdot \text{RELU}(\boldsymbol{w}_1 \cdot S))。 \tag{5}$$

式中:$\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 为 2 个权重矩阵,分别用于降维与升维,用于计算通道的重要性。$\boldsymbol{w}_1$ 先进行降维,减少计算量并提取通道特征。$\boldsymbol{w}_2$ 用于升维,恢复通道权重的原始形状。RELU 为激活函数,σ 为 Sigmoid 激活函数,用于限制通道权重在[0,1]之间,确保权重可学习。

(3)通道重新加权。通道重新加权公式如式(6)所示,最后计算出通道权重 $\boldsymbol{w}$,用于重新调整输

入特征图权重比例,提升重要通道权重。

$$F' = F \cdot w。 \tag{6}$$

式中:F 为原特征图;F' 为通过注意力增强后的特征图,F' 中的重要通道已被增强,冗余通道被抑制。

2.3 网络优化与模块应用

YOLOv12[16]作为当前 YOLO 系列性能最新的检测网络,在红外动态目标检测中,一方面 YOLOv12 仍以单帧图像作为输入进行检测,未考虑目标的运动信息,导致误检和漏检指标上升;另一方面在通道处理上通常使用 1×1 卷积进行固定权重的通道变换,无法根据输入数据的变化动态调整各个通道的重要性。在接收来自输入端 DAM 模块传入的通道融合特征后,由于在不同通道上的信息贡献不同,如果所有通道都被均等处理,会导致关键通道的贡献被稀释,影响检测效果。

为解决上述问题,本文以 YOLOv12s[16]作为基础网络,提出了一种基于 DAM 和 CACONV 的优化架构 YOLO-IDOD,使 YOLOv12 具备更强的红外动态目标检测能力,并提高其在红外动态目标检测任务中的泛化性,YOLO-IDOD 的网络结构如图 5 所示。

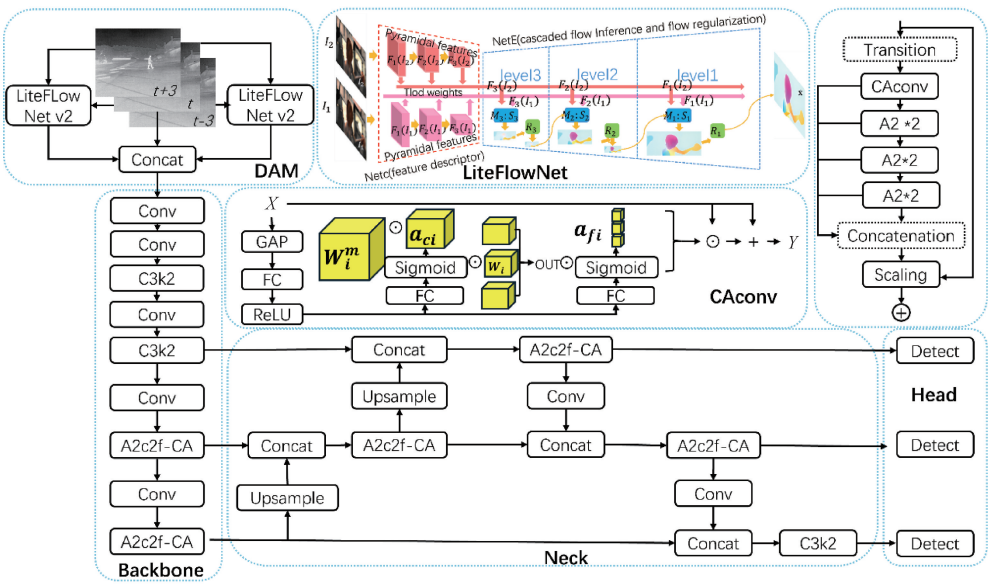


图 5 YOLO-IDOD 网络架构

Figure 5 YOLO-IDOD network architecture

在 YOLO-IDOD 中,DAM 主要用于优化输入端的数据特征,使得模型能够利用光流信息关注动态目标,同时抑制背景运动噪声。DAM 在输入端选取连续间隔 3 帧的原始帧进行光流计算并将结果进行通道融合,将动态目标的外观信息与运动信息进行融合。且在输入端引入 1×1 Conv-CA(Conv-CACONV)模块进行通道优化,该模块在 Backbone 输入端引入通道注意力模块,使网络可以针对输入的外观信息与运动信息进行权重学习。

CACONV 主要用于优化 Backbone 和 Neck 结构,使检测器更加关注关键特征通道,提高通道信息的有效性。在基础网络中,通过在 A2c2f[16]的 1×1 卷积模块中添加通道注意力机制从而生成 A2c2f-CA 变体,由于 A2c2f 本身的 A^2 模块拥有空间注意力机制,CA 变体可以让 A2c2f[16]在添加较少参数的同时实现多维注意力。

3 实验部分

3.1 数据集与评价指标

为证明改进算法的有效性,采用开源的公共数据集 FLIR_ADAS_v2[27]的部分连续稳定画面帧和自行拍摄的红外视频数据集 IRDA 进行实验,训练集和验证集配置如表 1 所示。

表 1 数据集设置

Table 1 Dataset settings

数据库	动态类	总数据	训练集	验证集
FLIR_ADAS_v2	12	3 749	3 000	749
IRDA	4	23 781	19 024	4 757

FLIR_ADAS_v2[27]数据集共包含 3 749 张连续稳定帧画面,提供 COCO 格式标注文件,图片尺寸为 640×512,包含人物、自行车、轿车、摩托车、公共汽车、火车、卡车、狗等共计 12 个类别为动态目标类别,主要目标实物图如图 6 所示。

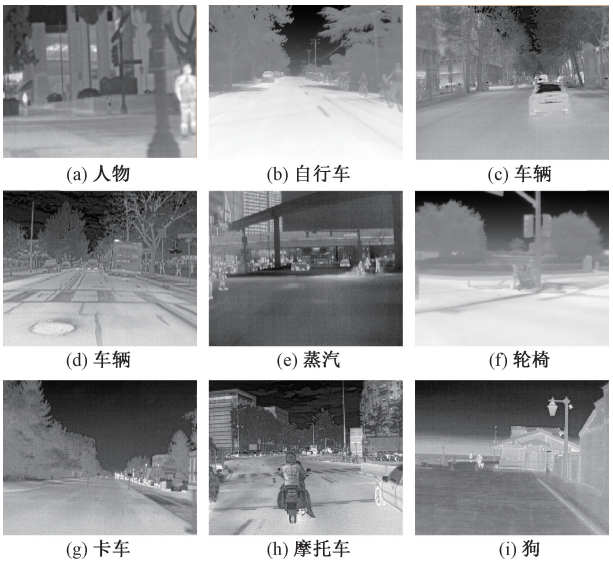


图 6　主要动态目标

Figure 6　Main dynamic objectives

IRDA 数据集为自建数据集,共包含视频场景 56 个,视频 332 个,画面帧数据集共 23 781 张,专业测试视频场景 10 个。图片尺寸为 640×512,包含烟头、人物、轿车、电瓶车、气体共计 5 类动态目标类别,其中气体作为非实质对象不作为本次检测目标,FLIR_ADAS_v2 数据集中人物与轿车类别与 IRDA 进行同类别合并处理,数据集总计包含 14 类动态检测目标。

测试集为验证模型实时检测效率与检测精度,通过裁剪画面稳定的 FLIR_ADAS_v2 公开源视频与 IRDA 源视频组成测试数据集。测试集要求视频场景

与训练集验证集互不重复,每段视频长度 30 s,每秒包含 25 帧图像,共 10 段视频,且场景相互独立,涵盖训练集中 14 类检测目标,为保障实时检测效果,单段视频模型检测时间不得超过 30 s。

本研究中模型性能评价指标采用准确率 P、召回率 R、平均准确率 mAP。并将推理时间 T 作为合格指标用以要求模型保持实时检测能力。其中准确率 P 为预测为正样本的结果中真实正样本的占有比例,反映模型预测结果的精确性;召回率 R 表示真实正样本中被模型检测为正样本所占的比例,反映模型对目标的检出能力;mAP 表示所有类别的 P-R 曲线下面积 AP 的平均值,用于评估模型在不同类别、不同阈值下的整体精确性与召回能力,其中 $mAP@50$ 代表 IOU 阈值固定为 0.5 时的平均值,$mAP@95$ 代表 IOU 阈值固定为 0.95 时的平均值;推理时间 T 用于衡量模型对单幅图像完成一次前向推理所需的平均时间,本次实验视频流每秒帧数为 25 帧,单帧平均推理时间低于 40 ms 则代表模型具备实时检测能力。

3.2　实验配置

本研究在统一的实验平台与参数设置下,对不同目标检测模型进行了对比实验。实验平台采用统一训练环境,具体配置如下:CPU 为 AMD Ryzen 9 5950X,16 核 32 线程,主频 3.4 GHz;GPU 为双 NVIDIA RTX 4090 显卡(各显存 24 GB,累计 48 GB),使用 CUDA12.1 与 cuDNN8.9;内存为 128 GB;操作系统为 Ubuntu22.04LTS。在训练参数方面,为保证公平对比,所有模型均在相同的数据集上训练 500 个 *Epoch*,采用默认 SGD 优化器[28],动量设置为 0.937,权重衰减为 0.000 5,预热学习率设置为 0.01,学习率调度器使用 cosine 衰减策略[12]。训练过程中 *batch_size* 设置为 64,输入图片尺寸设置为 640×640,开启 AMP 混合训练,使用包括 Mosaic、MixUp、HSV 随机增强、随机翻转等图像增强策略[12]。

3.3　消融实验与超参数分析

在 DAM 内部进行动态特征提取时,输入进光流网络的运作帧对间隔差距对动态信息的提取性能存在关键性影响,为了改善该超参数对于网络性能的影响。进行如下对比实验,将常规视频流帧率 25 帧作为标准帧率,以 YOLOv12s[16] 作为基准网络,将自建数据集(IRDA)中的动态目标作为检测对象。分别测试帧差参数为 1、3、5、7 帧时送入 DAM 模块的数据对于基础网络的目标检测性能影响,结果如表 2 所示。在时间差帧为 3 帧、时间差帧在 120 ms 时网络对于 DAM 提供的信息捕捉最为精确,其在 $mAP@50$ 和 $mAP@95$ 上相较于时间差帧为 1 帧、5 帧、7 帧均有领先,这代表时间差帧为 1 帧时物体运动信息较弱,光流网络无法提供足够强的动态信息,而在时间差帧为 5 帧与 7 帧时物体前后位移像素过大,检测网络无法准确提取 DAM 模块提供的动态特征。

表 2　DAM 超参数对比实验

Table 2　Hyperparameter comparison experiment

时差/ms	帧差/帧	精确率/%	召回率/%	$mAP50$/%	$mAP95$/%
40	1	74.8	66.8	72.5	55.7
120	3	**77.2**	**69.5**	**75.1**	**57.1**
200	5	71.7	63.2	71.1	52.2
280	7	65.1	44.3	62.2	41.8

为验证改进算法每一个模块的效果准确性,本文在测试数据集上进行消融实验,该数据集包括 14 个目标类别,可有效验证算法的准确性。

消融实验结果如表 3 所示,DAM 与 CACONV 模块在引入后均带来了不同程度的检测性能提升,表明两者对红外图像特征建模具有显著增强作用。具体来看,DAM 模块主要提升了召回率和整体检测稳定性,而 CACONV 模块在精细判别上表现更优。将两者结合形成的 YOLO-IDOD 模型在 *mAP*@50 与 *mAP*@95 指标上分别提升 4.6 百分点和 2.4 百分点,验证了模块设计的有效性与可组合性,尽管推理时间有所增加,但仍保持较好的实时性能,在提高了网络部分计算量的同时单帧图像处理时长仍维持在 40 ms 以内,符合实时检测要求,具备在实际红外检测场景中部署的潜力。

表 3 消融实验结果

Table 3 Results of ablation experiment

单位:%

DAM	CACONV	精确率	召回率	*mAP*@50	*mAP*@95
		74.3	62.2	73.1	54.9
√		77.2	**69.5**	75.1	57.1
	√	76.6	61.7	75.3	**57.6**
√	√	**79.9**	62.5	**77.7**	57.3

表 3 中消融实验结果显示:YOLO-DAM 模块相比 YOLOv12s[16] 基础模型显著提高了召回率,即从 62.2%提升至 69.5%,该现象表明 DAM 模块对降低漏检现象有显著贡献。而在引入 CACONV 模块后,虽然召回率出现了一定下降,但精确率和 *mAP* 指标总体进一步提升,模型总体性能改善。为验证 DAM 与 CACONV 对于模型漏检率与误检率的影响,进行如下实验设计,其中测试集中共 7 500 张测试图片,目标实例为 11 247 个。

表 4 为误检率分析,由表 4 可知,召回率明显提高至 69.5%,正确检测数量提升至 7 819 个,但伴随的误检数量较基准略有增加。单独使用 CACONV 模块时召回率有所下降,但误检数量减少,体现了其对背景误检的抑制效果。YOLO-IDOD 综合使用 DAM 与 CACONV 模块后,召回率稳定在 62.5%,误检数量明显降低至 1 770 个,表现出精确率的显著提升,体现出模块组合的优势。

表 4 误检率分析

Table 4 False positive rate analysis

DAM	CACONV	精确率/%	召回率/%	正检	误检	漏检
		74.3	62.2	6 998	2 420	4 251
√		77.2	**69.5**	**7 819**	2 310	**3 430**
	√	76.6	61.7	6 941	2 121	4 308
√	√	**79.9**	62.5	7 031	**1 770**	4 218

当仅接入 DAM 模块时,DAM 模块将光流运动信息与图像数据融合在输入端,模型在输入侧可以获得稳定的短时光流先验,使网络对于运动区域响应增强。当目标存在较明显运动特征时,光流特征突出明显,模型会大幅增强对目标区域的关注度。此时网络会趋向关注所有明显运动的区域,用以提高对动态目标的敏感性,导致召回率上升,但因此对于背景伪运动也会出现响应倾向:如树叶晃动、杂物漂浮等,从而导致误检率增加。

当仅接入 CACONV 模块时,网络在通道变换处引入可学习的通道重新加权,削弱冗余通道与噪声特征,但对于微弱运动和尺寸较小目标的响应被温和抑制,从而导致召回率下降。

而将 DAM 模块与 CACONV 模块进行结合引入,使模型兼具运动先验与通道判别能力,模型通过 CACONV 模块压制并过滤前端 DAM 模块提供的过于敏感的运动特征,同时保留动态目标的运动先验知识,在提高对于动态目标检测性能的同时抑制误检率。

3.4 算法对比实验与泛化性检测

为进一步验证改进算法的性能,本文在测试数据集上比较了 YOLO-IDOD 和其他实时检测算法检测性能的表现,通过将 YOLO-IDOD 与 YOLOv11[29]、YOLOv12[16]、RT-DETR[22]、RT-DETRv2[23] 和添加了多尺度卷积模块(multi-scale,MS)的 YOLOv12-MS[30] 进行比较,主要检测目标对象为连续帧下的红外动态目标。其中 RT-DETR[22] 系列网络是首个主流的以注意力机制为核心的端到端实时检测模型,相比传统检测器,RT-DETR[22] 通过引入高效的特征选择模块和多尺度特征融合机制,显著降低了计算复杂度,在保持精度的同时提升了推理速度;其改进版本 RT-DETRv2[23] 进一步优化了特征融合路径和查询解码策略,使网络更适用于实时场景下的小目标检测任务。而多尺度卷积模块则旨在增强网络的多尺度表征能力,该模块通过将输入特征图划分为多个通道组,并结合异构卷积核选择(heterogeneous kernel selection,HKS)策略,提升模型对尺度变化显著目标的感知能力,通过插入此模块,YOLO 网络可有效提高对于动态目标的检测精度与鲁棒性。

在表 5 中可以看出,与轻量级基准网络相比,本文所提出的 YOLO-IDOD 在保证运行速度的前提下,以红外图像中的动态对象为检测目标:*mAP*@50 提升了 4.6%,*mAP*@95 提升了 2.4%,精确率提升了 5.6%,召回率提升了 0.3%。

表 6 展示与其他算法的推理时间、计算量、参数量的数值比较,其中本文算法相较于基础网络在添

表 5　不同算法比较结果

Table 5　Comparison of the results of different algorithms

单位:%

算法	精确率	召回率	*mAP*50	*mAP*95
YOLOv11-S[29]	71.3	62.9	72.1	53.8
YOLOv12-S[16]	74.3	62.2	73.1	54.9
RT-DETR-R50[22]	68.3	57.4	65.7	49.5
RT-DETRv2-R50[23]	69.4	59.9	66.0	49.9
YOLOv12-MS-S[30]	74.7	**63.0**	73.0	54.8
YOLO-IDOD-S	**79.9**	62.5	**77.7**	**57.3**

加 DAM 与 CACONV 模块情况下,参数量与计算量仅小幅增加,且模型单帧延迟 T 仍满足实时阈值 $T<40$ ms。

表 6　不同算法复杂度与推理时间比较

Table 6　Comparison of the computational complexity and inference time of different algorithms

算法	参数量/10^6	计算量/GFLOPs	推理时间/ms
YOLOv11-S[29]	9.4	21.5	**3.3**
YOLOv12-S[16]	9.3	**21.4**	3.6
RT-DETR-R50[22]	42.0	136.0	10.8
RT-DETRv2-R50[23]	42.0	136.0	10.0
YOLOv12-MS-S[30]	**8.4**	21.4	4.0
YOLO-IDOD-S	12.1	29	21.7

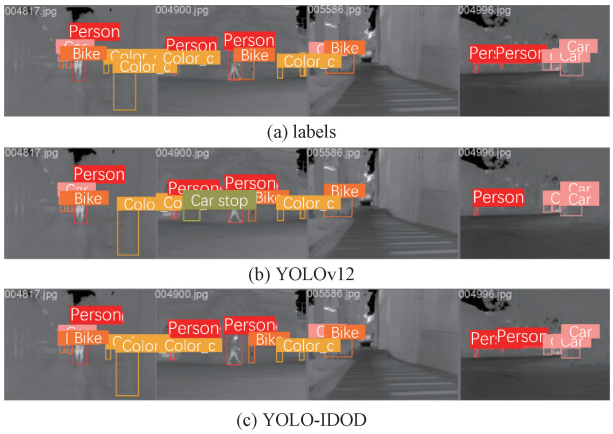
综合表 5、表 6 数据分析可得,YOLO-IDOD 在不显著增加推理时间的前提下换取了对红外动态目标更稳健的识别,满足实时性的同时相较于各类基础算法提升了较大精度。

为验证改进算法的泛化能力,在 IRVI[31] 数据集上进行泛化对比实验。YOLO-IDOD 与 YOLOv12[16] 在 IRVI[31] 数据集上的预测效果如图 7 所示。其中 YOLO-IDOD 相较于 YOLOv12 检测结果具有如下优势:

(1)检测精度提升显著:在图像 004 817.jpg 与 004 996.jpg 中,YOLOv12[16] 对于 color_cone 与 car 类目标存在明显漏检,而 YOLO-IDOD 能完整检测所有类别目标;图像 004 900.jpg 中,YOLOv12 仅检测出部分 person 与 bike,而 YOLO-IDOD 能准确识别全部目标,并给出更高置信度评分。

(2)误检率降低:在图像 004900.jpg 中,YOLOv12 错误识别了“car_stop”类别目标(绿色框),属于误检;而 YOLO-IDOD 能有效抑制此类误检现象,提升分类判别鲁棒性。

(3)边界框定位更精准:YOLO-IDOD 的框体与目标实际边界更加贴合,尤其是在车类、person 类小目标上体现更明显。



(a) labels

(b) YOLOv12

(c) YOLO-IDOD

图 7　不同算法在 IRVI[31] 的比较结果

Figure 7　Comparison of the results of different algorithms on the IRVI[31]

YOLO-IDOD 之所以在泛化能力和检测精度上表现优异,源于其结构上的两大改进模块:

(1)DAM 通过捕捉红外视频序列中目标的时序变化特征增强动态信息表达,提高了对低对比度目标的辨识能力,解决了 YOLOv12 在静态帧中因动态目标模糊而导致的漏检问题。

(2)CACONV 模块在特征提取阶段引入通道注意机制,有助于模型聚焦于红外图像中的高响应区域,从而提升小目标与边界区域的识别效果,抑制背景噪声误导,降低误检风险。

实验结果如表 7 所示,表内结果表明改进后的 YOLO-IDOD 算法在保证运行速度的前提下在红外动态目标检测任务中相较于基础网络拥有更高的检测精度和泛化性。

表 7　在 IRVI[31] 上的泛化性测试

Table 7　Generalization test on IRVI[31]

单位:%

算法	精确率	召回率	*mAP*@50	*mAP*@95
YOLOv12[16]	63.3	55.7	65.3	43.8
YOLO-IDOD	**66.9**	**62.2**	**69.1**	**44.2**

4　结论

针对现有红外图像中对动态目标的检测存在精度不高的问题,本文提出了一种 YOLO-IDOD 算法.该算法以 YOLOv12[16] 作为基础模型,通过嵌入 DAM 实现时序信息嵌入使模型更关注运动目标,通过嵌入 CACONV 实现通道注意力机制,提升通道信息利用率,提升了红外动态目标检测的准确性和鲁棒性。实验结果验证了该方法的有效性,在提升检测精度的同时仍保持实时的推理速度,同时上述模块也可作为优化动态目标检测模型的即插即用模块在更广泛的目标检测领域上受到应用。

但当前 YOLO-IDOD 仅在高性能主机上才能获得实时性保证,一方面,如何优化 YOLO-IDOD 的网络结构并进一步提升推理速度仍然是一个值得深度研究的课题。另一方面,动态目标中红外气体由于其本身非实体性质导致光流网络无法捕捉其动态特征,下一步我们将尝试实现对红外气体的检测能力和提升红外动态目标整体的推理速度,以提高方法的普适性。

参考文献:

[1] Xu Huilin, Zhao Xin, Yu Bo, et al. Multi-resolution feature extraction algorithm for semantic segmentation of infrared images[J]. Infrared Technology, 2024, 46(5): 556-564. [徐慧琳, 赵鑫, 于波, 等. 一种多分辨率特征提取红外图像语义分割算法[J]. 红外技术, 2024, 46(5): 556-564.]

[2] Li Yuanbo, Zhou Ping, Zhou Gongbo, et al. A comprehensive survey of visible and infrared imaging in complex environments: principle, degradation and enhancement [J]. Information Fusion, 2025, 119: 103036.

[3] Chen Tianxiang, Ye Zi, Tan Zhentao, et al. MiM-ISTD: mamba-in-mamba for efficient infrared small-target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5007613.

[4] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2014-09-01)[2025-11-10]. https://doi.org/10.48550/arXiv.1409.0473.

[5] Ye Baicheng, Zhu Youpan, Zhou Yongkang, et al. Review of lightweight target detection algorithms[J]. Infrared Technology, 2025, 47(3): 289-298. [叶栢铖, 朱尤攀, 周永康, 等. 轻量级目标检测算法综述[J]. 红外技术, 2025, 47(3): 289-298.]

[6] Guo Haofan, Jiao Ting, Sun Fangliang, et al. Real-time infrared imaging gas-leak detection method based on improved YOLOv5-seg[J]. Infrared Technology, 2025, 47(7): 918-927. [郭浩帆, 焦婷, 孙方亮, 等. 基于改进 YOLOv5-Seg 的实时红外成像气体泄漏检测方法[J]. 红外技术, 2025, 47(7): 918-927.]

[7] Dai Yimian, Wu Yiquan, Zhou Fei, et al. Attentional local contrast networks for infrared small target detection [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(11): 9813-9824.

[8] Yue Taoran, Lu Xiaojin, Cai Jiaxi, et al. YOLO-MST: multiscale deep learning method for infrared small target detection based on super-resolution and YOLO[J]. Optics & Laser Technology, 2025, 187: 112835.

[9] Wang Quan, Liu Fengyuan, Cao Yi, et al. LFIR-YOLO: lightweight model for infrared vehicle and pedestrian detection[J]. Sensors, 2024, 24(20): - .

[10] Sun Mingyuan, Zhang Haochun, Huang Ziliang, et al. Road infrared target detection with I-YOLO[J]. IET Image Processing, 2022, 16(1): 92-101.

[11] Ling Song, Hong Xianggong, Liu Yongchao. YOLO-APDM: improved YOLOv8 for road target detection in infrared images[J]. Sensors, 2024, 24(22): 7197.

[12] Sohan M, Sai Ram T, Rami Reddy C V. A review on YOLOv8 and its advancements[C]//Data intelligence and cognitive informatics. Singapore: Springer Nature Singapore, 2024: 529-545.

[13] Wang Yong, Wang Bairong, Huo Lile, et al. GT-YOLO: nearshore infrared ship detection based on infrared images [J]. Journal of Marine Science and Engineering, 2024, 12(2): 213.

[14] Zhao Xiaofeng, Zhang Wenwen, Zhang Hui, et al. ITD-YOLOv8: an infrared target detection model based on YOLOv8 for unmanned aerial vehicles [J]. Drones, 2024, 8(4): 161.

[15] Hao Xinyue, Luo Shaojuan, Chen Meiyun, et al. Infrared small target detection with super-resolution and YOLO [J]. Optics & Laser Technology, 2024, 177: 111221.

[16] Tian Yunjie, Ye Qixiang, Doermann D. YOLOv12: attention-centric real-time object detectors[PP/OL]. V1. arXiv (2025-02-18)[2025-11-10]. https://doi.org/10.48550/arXiv.2502.12524.

[17] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 779-788.

[18] Liu Wei, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//Computer vision - ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[19] Chen Yuming, Yuan Xinbin, Wang Jiabao, et al. YOLO-MS: rethinking multi-scale representation learning for real-time object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4240-4252.

[20] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[PP/OL]. V7. arXiv (2023-08-02)[2025-11-10]. https://doi.org/10.48550/arXiv.1706.03762.

[22] Zhao Yian, Lv Wenyu, Xu Shangliang, et al. DETRs beat YOLOs on real-time object detection[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 16965-16974.

[23] Lv Wenyu, Zhao Yian, Chang Qinyao, et al. RT-DETRv2: improved baseline with bag-of-freebies for real-time detection transformer[PP/OL]. (2024-07-24)[2025-11-10]. https://doi.org/10.48550/arXiv.2407.17140.

[24] Hui T W, Tang Xiaoou, Loy C C. A lightweight optical flow CNN—revisiting data fidelity and regularization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2555-2569.

[25] Lin Min, Chen Qiang, Yan Shuicheng. Network in network[PP/OL]. (2014-03-04)[2025-11-10]. https://doi.org/10.48550/arXiv.1312.4400.

[26] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]//Proceedings of the 2018 IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.
[27] Uzun E, Dursun A A, Akagündüz E. Augmentation of atmospheric turbulence effects on thermal adapted object detection models[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2022: 240-247.
[28] Gower R M, Loizou N, Qian Xun, et al. SGD: general analysis and improved rates[PP/OL]. V4. arXiv (2019-05-01)[2025-11-10]. https://doi.org/10.48550/arXiv.1901.09401.
[29] Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements[PP/OL]. (2024-10-23)[2025-11-10]. https://doi.org/10.48550/arXiv.2410.17725.
[30] Chen Yuming, Yuan Xinbin, Wang Jiabao, et al. YOLO-MS: rethinking multi-scale representation learning for real-time object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4240-4252.
[31] Li Shuang, Han Bingfeng, Yu Zhenjie, et al. I2V-GAN: unpaired infrared-to-visible video translation[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 3061-3069.

Real-time Detection Algorithm for Infrared Dynamic Targets Based on YOLO-IDOD

ZHAO Xin[1,2], FEI Xiaohu[1], WANG Dongyu[1], HAN Shoufei[1]

(1. School of Artificial Intelligence, Anhui University of Science and Technology, Anhui, Huainan 232001, China; 2. The development of intelligent technology for the mechanised extraction of coal is being conducted at the National Key Laboratory for Numerical Simulation of Geomechanics, Chinese Academy of Sciences, Anhui, Huainan 232001, China)

Abstract: To overcome the limitation that existing infrared object detection algorithms had inadequately exploited temporal information and inter-frame dependencies in dynamic target detection, thereby resulting in suboptimal detection accuracy, a real-time infrared dynamic object detection framework based on YOLO-IDOD, incorporating a Dynamic Attention Module (DAM) and a Channel Attention Convolution (CACONV) module, had been proposed. The YOLOv12s architecture had been employed as the baseline network, in which a dynamic attention mechanism had been integrated at the input stage to extract short-term optical flow features via an optical flow network, effectively suppressing background motion interference and enhancing the network's sensitivity to target motion characteristics. Furthermore, a channel attention convolution module had been embedded within the network architecture, where channel-wise attention mechanisms had been introduced at both the input and output stages to facilitate more discriminative feature representation and selection for the DAM-enhanced features. The proposed modules had been designed as plug-and-play components, enabling spatiotemporal feature aggregation and adaptive feature selection, thereby improving the generalization capability of the network for infrared dynamic target detection. Experimental evaluations had demonstrated that the improved YOLO-IDOD model had achieved a precision of 79.9%, a recall of 62.5%, an mAP@50 of 77.7%, and an mAP@95 of 57.3% on a mixed dataset composed of a self-constructed dataset (IRDA) and the public FLIR_ADAS_v2 dataset. Compared with the baseline YOLOv12s model, precision, mAP@50, and mAP@95 had been improved by 5.2, 4.6, and 2.4 percentage points, respectively, while maintaining a comparable recall rate, thereby effectively enhancing detection accuracy and generalization performance for infrared dynamic targets.

Keywords: Infrared Dynamic Target Detection; YOLOv12; DAM; CACONV; Multi-dimensional channel attention mechanism